Simran Arora (https://arorasimran.com/)

Overview

My research is in **AI Systems**. Using techniques in systems and in ML, work on developing AI systems that scale and use the hardware efficiently. Three fundamental questions I study are:

- **Hardware** As AI hardware grows in complexity, how do we make it easier to extract high utilization?
- **Software** How do we build AI algorithms that scale and use the hardware efficiently?
- **Applications** What are the most efficient ways to use our models in AI applications?

Education

2019–	PhD in Computer Science, Stanford University Research advisor: Christopher Ré
2015-2019	Jerome Fisher Management & Technology Program, University of Pennsylvania *Summa cum laude BSE in Computer Science, School of Engineering* *Summa cum laude BS in Economics (Finance Concentration), The Wharton School* *Minor in Mathematics* Research advisors: Boon Thau Loo, Vijay Kumar, Vincent Liu

Experience

2019-	PhD Student, **Stanford University**
2024-	Academic Partner, **Together AI**
2023-	Advisor, **Cartesia AI**
2013-	Academic Partner, **Looma Education, AI**
2021-2022	Research Scientist, **Facebook AI Research** Worked with Jacob Kahn, Patrick Lewis, Angela Fan, and Ronan Collobert Published our research at TACL
Summer 2018	Technology Investment Banking Intern, **Morgan Stanley, Menlo Park** Worked on the sale of Acxiom AMS to IPG for $2.3Bn, sale of Cylance to Blackberry for $ 1.7Bn, and Sonos IPO on Nasdaq
Summer 2017	Software Engineer, **Google**

Awards

July 2024	**ICML ES-FoMo Best Paper Award (Amongst 83 Papers)** Simple linear attention language models balance the recall-throughput tradeoff
July 2024	**ICML Spotlight Award (Top 3.5% of 10K Submitted Papers)** Simple linear attention language models balance the recall-throughput tradeoff
December 2023	**NeurIPS Outstanding Paper Award (4 Papers in 12.3K Submissions)** DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models
December 2023	**NeurIPS Oral Award (Top 0.5% of Papers)** Monarch Mixer: A Simple Sub-Quadratic GEMM-Based Architecture
April 2023	**ICLR Spotlight Award (Top 25% of Accepted Papers)** Ask Me Anything: A simple strategy for prompting language models
February 2023	**AAAI KnowledgeNLP Workshop: Oral Award (Top 20% of Papers)** Reasoning over Public and Private Data in Retrieval-Based Systems
2019-2023	**Stanford Graduate Fellowship, 3 years**
2019	**Rhodes Scholarship National Finalist**
2019	**Marshall Scholarship National Finalist**

2019	**Penn Computer Science Academic Award**
	One graduating CS major per year
2019	**Michele Huber and Bryan D. Giles Memorial Award**
	One graduating Jerome Fisher student per year
2019	**Penn Computer Science Senior Engineering Capstone Project 2^{nd} Place**
2019	**Wharton School Summa Cum Laude (Highest Honors)**
2019	**Penn Engineering Summa Cum Laude (Highest Honors)**
July 2017	**Best Paper Runner Up: IEEE MARSS Conference**
	Control of multiple microrobots with multiscale magnetic field superposition
2017-2019	**University of Pennsylvania Tau Beta Pi and Eta Kappa Nu**
2015	**International University Physics Competition, Silver Medalist**
2015	**Siemens Research Competition Semifinalist**

Selected Research

[1] Benjamin Spector, Simran Arora, Aaryan Singhal, Daniel Fu, and Christopher Ré
ThunderKittens: Simple, Fast, and Adorable Kernels, In Submission.

[2] Michael Zhang, Simran Arora, Rahul Chalamala, Benjamin Spector, Alan Wu, Krithik Ramesh, Aaryan Singhal, and Christopher Ré
LoLCATS: Low-rank Linearization of Large Language Models, In Submission and ICML ES-FoMo 2024.

[3] Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, Silas Alberti, James Zou, Atri Rudra, and Christopher Ré
Simple linear attention language models balance the recall-throughput tradeoff
International Conference on Machine Learning (ICML), 2024. **Spotlight Award** and ICML ES-FoMo, 2024. **Oral Award, Best Paper Award**

[4] Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré
Zoology: Measuring and Improving Recall in Efficient Language Models
International Conference on Learning Representations (ICLR), 2024.

[5] Simran Arora, Avanika Narayan, Mayee F. Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Ré
Ask Me Anything: A simple strategy for prompting language models
International Conference on Learning Representations (ICLR), 2023. **Spotlight Award**

Select Open Source Artifacts

- **Systems:** ThunderKittens GitHub (1.6K+ stars), Evaporate GitHub (500 stars), Bootleg GitHub (200+ stars), ConcurrentQA GitHub
- **Efficient architectures and algorithms:** Based GitHub (200+ stars), LoLCATS GitHub (200 stars), Long-context BERT models: Monarch Mixer GitHub and M2-BERT Retrieval Model Checkpoints (500+ stars), Ask Me Anything GitHub (500+ stars)

Select Industry Use and Press

BASED and LoLCATS Together AI, Monarch Mixer Together AI, Mongo DB, LangChain, LlamaIndex and Nomic AI; Bootleg Apple ML Research; ConcurrentQA Meta AI Research; Privacy Venture Beat; Evaporate LlamaIndex; Ask Me Anything Snorkel AI / Numbers Station; Data Wrangling Numbers Station

Publications

Systems Work

[1] Benjamin Spector, Simran Arora, Aaryan Singhal, Daniel Fu, and Christopher Ré
ThunderKittens: Simple, Fast, and Adorable Kernels, In Submission.

[2] Megha Srivastava, Simran Arora, and Dan Boneh
Optimistic Verifiable Training by Controlling Hardware Nondeterminism,
Advances in Neural Information Processing Systems (NeurIPS), 2024 and ICML ES-FoMo 2024.

[3] Furui Cheng, Vilém Zouhar, Simran Arora, Mrinmaya Sachan, Hendrik Strobelt, and Mennatallah El-Assady
RELIC: Investigating Large Language Model Responses using Self-Consistency
Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI), 2024.

[4] Simran Arora, Brandon Yang, Sabri Eyuboglu, Avanika Narayan, Andrew Hojel, Immanuel Trummer, and Christopher Ré
Language Models Enable Simple Systems for Generating Structured Views of Heterogeneous Data Lakes
Proceedings of the VLDB Endowment (PVLDB), 2023.

[5] Simran Arora, Patrick Lewis, Angela Fan, Jacob Kahn, and Christopher Ré
Reasoning over Public and Private Data in Retrieval-Based Systems
Transactions of the Association for Computational Linguistics (TACL), 2023 and AAAI KnowledgeNLP, 2023. **Oral Award**

[6] Avanika Narayan, Laurel Orr, Ines Chami, Simran Arora, and Christopher Ré
Can Foundation Models Wrangle Your Data?
Proceedings of the VLDB Endowment (PVLDB), 2022.

[7] Laurel Orr, Megan Leszczynski, Simran Arora, Sen Wu, Neel Guha, Xiao Ling, and Christopher Re
Bootleg: Chasing the Tail with Self-Supervised Named Entity Disambiguation
Conference on Innovative Data Systems Research (CIDR), 2021.

[8] Qizhen Zhang, Akash Acharya, Hongzhi Chen, Simran Arora, Ang Chen, Vincent Liu, Boon Thau Loo
Optimizing Declarative Graph Queries at Large Scale
Proceedings of the 2019 International Conference on Management of Data (SIGMOD), 2019.

AI Work

[9] Michael Zhang, Simran Arora, Rahul Chalamala, Benjamin Spector, Alan Wu, Krithik Ramesh, Aaryan Singhal, and Christopher Ré
LoLCATS: Low-rank Linearization of Large Language Models, In Submission and ICML ES-FoMo 2024.

[10] Simran Arora, Aman Timalsina, Aaryan Singhal, Benjamin Spector, Sabri Eyuboglu, Xinyi Zhao, Ashish Rao, Atri Rudra, and Christopher Ré
Just read twice: closing the recall gap for recurrent language models, ICML ES-FoMo 2024.

[11] Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, Silas Alberti, James Zou, Atri Rudra, and Christopher Ré
Simple linear attention language models balance the recall-throughput tradeoff
International Conference on Machine Learning (ICML), 2024. **Spotlight Award** and ICML ES-FoMo, 2024. **Oral Award, Best Paper Award**

[12] Jon Saad-Falcon, Daniel Y. Fu, Simran Arora, Neel Guha, and Christopher Ré
Benchmarking and Building Long-Context Retrieval Models with LoCo and M2-BERT
International Conference on Machine Learning (ICML), 2024 and ICML ES-FoMo, 2024.

[13] Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré
Zoology: Measuring and Improving Recall in Efficient Language Models
International Conference on Learning Representations (ICLR), 2024.

[14] Jerry Liu, Jessica Grogan, Owen Dugan, Simran Arora, Atri Rudra, and Christopher Ré
Can Transformers Solve Least Squares to High Precision?, ICML ES-FoMo 2024.

[15] Sabri Eyuboglu, Dylan Zinsley, Jon Saad-Falcon, Simran Arora, Atri Rudra, James Zou, Chris Ré
Towards smaller language models via layer looping, ICML ES-FoMo 2024.

[16] Daniel Y. Fu, Simran Arora, Jessica Grogan, Isys Johnson, Sabri Eyuboglu, Armin W. Thomas, Benjamin F. Spector, Michael Poli, Atri Rudra, and Christopher Ré
Monarch Mixer: A Simple Sub-Quadratic GEMM-Based Architecture
Advances in Neural Information Processing Systems (NeurIPS), 2023. **Oral Award**

[17] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li
DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models
Advances in Neural Information Processing Systems (NeurIPS), 2023. **Oral Award** and **Outstanding Paper Award**

[18] Simran Arora, Avanika Narayan, Mayee F. Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Ré
Ask Me Anything: A simple strategy for prompting language models
International Conference on Learning Representations (ICLR), 2023. **Spotlight Award**

[19] Simran Arora and Christopher Ré
Can Foundation Models Help Us Achieve Perfect Secrecy?, AAAI PPAI Workshop, 2023.

[20] Simran Arora, Sen Wu, Enci Liu, and Christopher Re
Metadata shaping: A simple approach for knowledge-enhanced language models
Findings of the Association for Computational Linguistics (ACL), 2022.

[21] Simran Arora, Avner May, Jian Zhang, and Christopher Ré
Contextual Embeddings: When Are They Worth It?
Proceedings of the Association for Computational Linguistics (ACL), 2020.

Other Work

[22] Edward Steager, Denise Wong, Jeremy Wang, Simran Arora, and Vijay Kumar
Control of multiple microrobots with multiscale magnetic field superposition
International Conference on Manipulation, Automation and Robotics at Small Scales (MARSS), 2017. **Best Paper Runner Up**

[23] B. P. Mason, M. Whittaker, J. Hemmer, Simran Arora, A. Harper, S. Alnemrat, A. McEachen, S. Helmy, J. Read de Alaniz, and J. P. Hooper
A temperature-mapping molecular sensor for polyurethane-based elastomers
Applied Physics Letters (APL), 2016.

Invited Talks

Panel Discussions

Simon's Institute: Are Transformers the end game? (Berkeley, CA): Summer 2024
with Jitendra Malik, Stella Biderman, Andrew Gordon Wilson

Understanding and Improving Efficient Models

Simon's Institute: Transformers as a Computational Model (Berkeley, CA): Summer 2024
Stanford NLP Group (Stanford, CA): Summer 2024
UC Berkeley NLP Group (Berkeley, CA): Summer 2024
CCAIM Summer School (Virtual): Summer 2024
Liquid AI (Vienna, Austria): Summer 2024
Princeton University PLI Group (Princeton, NJ): Winter 2024
Cornell Tech (New York, NY): Winter 2024
Microsoft AI Research (Virtual): Winter 2024
56th Annual ACM Symposium on Theory of Computing (STOC) Workshop Keynote Speaker (Vancouver, Canada): Summer 2024

Building High Throughput Data Management Systems

NeurIPS TRL Workshop Keynote Speaker (New Orleans, LA): Winter 2023

Ask Me Anything: How are Foundation Models Changing the Way We Build Data Systems?

Snorkel Foundation Model Summit (Virtual): Winter 2023
Apple Machine Learning Research (Cupertino, CA): Winter 2023
Stanford CRFM Research Spotlight Talk (Stanford, CA): Fall 2023

Can Foundation Models Help Us Achieve Perfect Secrecy?

IBM AI Research (Virtual): Fall 2022
MIT CSS Seminar (Virtual): Spring 2023
Stanford HAI: AI and Society (Stanford, CA): Spring 2023
Oral at KnowledgeNLP-AAAI'23 (Washington DC): Winter 2023

Metadata Shaping: A Simple Approach for Knowledge-Enhanced Language Models

Facebook AI Research Reading Group (Virtual): Summer 2021
Spotlight at Stanford HAI Data-Centric AI Workshop (Virtual): Fall 2021

Contextual Embeddings: When are they worth it?

ACL Conference (Virtual): Summer 2020

Teaching

Fall 2023	**Course Co-Creator and Co-Instructor**, *CS 229S: Systems for Machine Learning* Stanford University, 3-Unit Undergrad-Graduate course. Taught 110+ students.
Fall 2023	**Instructor** *CS: 528: Machine Learning Systems Seminar* Stanford University
Spring 2019	**Course Co-Creator**, *MCIT 595: Computer Systems* University of Pennsylvania
Fall 2018	**Course Assistant**, *CIS 380: Operating Systems* University of Pennsylvania
Fall 2017 and Spring 2018	**Course Assistant**, *CIS 160: Discrete Mathematics* University of Pennsylvania

Educational Notes

- CS 229s Systems for ML course lecture notes
- Efficient architectures as arithmetic circuits blogpost
- Easier, better, faster, cuter blogpost
- Linearizing LLMs with LoLCATS blogpost
- GPUs Go Brrr blogpost
- Long-Context Retrieval Models with Monarch Mixer blogpost
- Just read twice: closing the recall gap for recurrent language models blogpost
- Based: Simple linear attention language models balance the recall-throughput tradeoff blogpost
- Zoology: Measuring and Improving Recall in Efficient Language Models blogpost
- Monarch Mixer: Revisiting BERT, Without Attention or MLPs blogpost
- The Safari of Deep Signal Processing: Hyena and Beyond blogpost

Mentorship

2024-	Aaryan Singhal (Stanford Undergrad, In submission ICLR 2025 paper)
2024-	Jerry Liu (Stanford CS PhD, First author paper at NeurIPS 2024 ES-FoMo. In submission paper ICLR 2025)
2023-2024	Xinyi ("Jojo") Zhao (Stanford CS MS, Snowflake DB)
2023-2024	Ashish Rao (Stanford CS Undergrad/Coterm, NVIDIA)
2023-2024	Jon Saad-Falcon (Stanford CS PhD, First author paper at ICML 2024)
2022-2023	Soumya Chatterjee (Stanford CS MS, First author paper at SIGIR REML 2023, now ML at Apple)
2022-2023	Andrew Hojel (Stanford CS Undergrad/Coterm, VLDB paper, now Member of the Technical Staff at Essential AI)
Fall 2022	Katie Giosio (Stanford CS PhD)
2021-2022	Enci Liu (Stanford CS Undergrad/Coterm, ACL paper, now ML at Apple)

Service

Reviewer	ICML (Top Reviewer Award), NeurIPS, ACL, PPAI-AAAI, NeurIPS TRL, ICLR ME-FoMo, ICML ES-FoMo
2023	Co-organized workshop on Decentralized and Collaborative Learning at MLSys
Summer 2023	Organized Stanford NLP Group weekly meetings
2022-2023	Stanford Center for Research on Foundation Models (CRFM) Leadership Team
2019-2022	Undergrad Mentor in the Stanford Women in STEM Mentorship Program
2018-2019	Undergrad Mentor in the Penn Women in CS Mentorship Program

Last updated: November 12, 2024 *

*CV template by Neel Guha, Daniel Fu, and Christopher Morris.